Faculty of Engineering and Information Technology
University of Technology, Sydney

Outlier Detection in Large High-Dimensional Data and Its Application in Stock Market Surveillance

A thesis submitted in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

by

Chao Luo

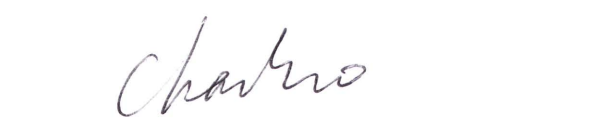
February 2011

CERTIFICATE OF AUTHORSHIP/ORIGINALITY

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Signature of Candidate



Acknowledgments

I would like to express my deep and sincere gratitude to my principal supervisor, Professor Chengqi Zhang, for helping me to get scholarship and offering me the chance to this study. Without his support, this thesis would be impossible. He has given me valuable guidance on my research during my doctoral study. He has provided opportunities for me to join industry projects which made me accumulate professional experience and knowledge.

I am deeply indebted to my co-supervisor Dr. Yanchang Zhao. The successful completion of this dissertation is due in great part to his support and guidance. I appreciate all his contributions of time, ideas and advices to help me complete this research. During the past three years, he has discussed with me regularly. He helped me develop ideas, design experiments, write academic papers and so on. I learned not only the skills of academic research, but also the right attitude on research from him. He taught me how to do interesting and meaningful research.

I am grateful for Prof. Longbing Cao for his suggestions and guidance on my research. He has provided me the good research environment and resources during my PhD study.

I am also grateful to my family for their continuous support in life. My parents, my brother and my sisters encouraged me and supported me whenever I met difficulty. Their love for me gave me strength to face difficulties over the years.

My additional thanks goes to the following schoolfellows for their suggestions and help during my study: Dr. Yuming Ou, Dr Jiarui Ni, Mr Zhigang

Zheng, etc., all members in the Data Sciences and Knowledge Discovery Research Lab, as well as all relevant staff and students in the Faculty and the University Graduate School. During my doctoral study, they gave me many help on study.

Finally, I appreciate the support from the CMCRC Scholarship for my research, which contributes to the delivery of this thesis.

Contents

List of Figures

List of Tables

Abstract

Outlier detection techniques play an important role in stock market surveillance that involves analysis of large volume of high-dimensional trading data. However, outlier detection in large high-dimensional data is very challenging and is not well addressed by existing techniques. Firstly, it is difficult to select useful and relevant features from high-dimensional data. Secondly, large high-dimensional data need more efficient algorithms.

To attack the above issues brought by large high-dimensional data, this thesis presents two outlier detection models and one subspace clustering model.

Firstly, an outlier mining model is proposed to detect the outliers from multiple complex stock market data. In order to improve the efficiency of outlier detection, a financial model is used to select the features to construct multiple datasets. This model is able to improve the precision of outlier mining on individual measurements. The experiments on real-world stock market data show that the proposed model is effective and outperforms traditional technologies.

Secondly, in order to find relevant features automatically, an agent-based algorithm is proposed to discover subspace clusters in high dimensional data. Each data object is represented by an agent, and the agents move from one local environment to another to find optimal clusters in subspaces. Heuristic rules and objective functions are defined to guide the movements of agents, so that similar agents(data objects) go to one group. The experimental results show that our proposed agent-based subspace clustering algorithm performs

better than existing subspace clustering methods on both F1 measure and Entropy. The running time of our algorithm is scalable with the size and dimensionality of data. Furthermore, an application of our technique to stock market surveillance demonstrates its effectiveness in real world applications.

Finally, we propose a reference-based outlier detection model by agent-based subspace clustering. At first, agent-based subspace clustering is utilized to generate clusters in subspaces. After that, the centers of clusters, together with the corresponding subspaces, are used as references, and a reference-based model is employed to find outliers in relevant subspaces. The experimental results on real-world datasets prove that the proposed model is able to effectively and efficiently identify outliers in subspaces.

In summary, this thesis research on outlier detection techniques on high-dimensional data and its application in stock market surveillance. The proposed models are novel and effective. They have shown their potentials in real business.